Please quote as: Tolzin, A., Knoth, N. & Janson, A. (2024). Leveraging Prompting Guides as Worked Examples for Advanced Prompt Engineering Strategies. International Conference on Information Systems (ICIS), Bangkok, Thailand.

Leveraging Prompting Guides as Worked Examples for Advanced Prompt Engineering Strategies

Completed Research Paper

Antonia Tolzin
Information Systems,
University of Kassel, Germany
antonia.tolzin@uni-kassel.de

Nils Knoth
Institute for Psychology,
University of Kassel, Germany
nils.knoth@uni-kassel.de

Andreas Janson
Institute of Information Systems and Digital Business
University of St. Gallen, Switzerland
andreas.janson@unisg.ch

Abstract

Artificial intelligence systems, particularly those based on large language models, are increasingly prevalent in personal and professional settings, making the skill of prompt engineering—formulating effective AI inputs—vital. This paper explores the effects and how to enhance students' prompt engineering skills based on a multi-study design. Our first study confirmed the baseline hypothesis that prompt engineering can predict AI output quality, framing it as a critical skill. We then investigated whether instructional designs (worked examples & instructions), could develop prompting skills. Using worked examples, the second study tested the effectiveness of instructional materials on enhancing prompting skills. The experiment involved 245 students who demonstrated that a brief exposure to a worked example-based prompting guide significantly improved their ability to deploy targeted prompting strategies. These findings suggest that integrating worked examples into curricula could effectively equip students with essential prompting skills, offering both theoretical and practical implications for AI education.

Keywords: Large Language Model, Worked Examples, Prompt Engineering, AI Interaction, Education

Introduction

In recent years, Artificial Intelligence (AI) has made significant progress in various fields, demonstrating advances in areas such as image recognition, speech understanding, and language processing (Berg et al., 2023). The emergence of cutting-edge technologies such as the ChatGPT language model holds immense promise for reshaping the educational landscape (Sok & Heng, 2023). Its remarkable ability to generate complex text swiftly distinguishes it from predecessors, raising concerns about student assessment, while LLMs and conversational interfaces increasingly enhance human communication and user experience (Dwivedi et al., 2023; McLean & Osei-Frimpong, 2019; Rudolph et al., 2023). LLMs simulate human language by predicting likely subsequent words, generating high-quality output that excites and concerns educators, with the potential to revolutionize tasks like essay writing (Bommasani et al., 2021; Choi et al., 2023; Dwivedi et al., 2023). Recent studies highlight its potential to improve learning through personalized and adaptive approaches (Rahman & Watanobe, 2023; Rasul et al., 2023; Zhu et al., 2023). Furthermore, given the increasing demand for skilled workers who are proficient in the use of these technologies, it is

imperative to train students in prompt engineering, as demonstrated by Dell'Acqua et al. (2023) in the context of consultant work outcomes, indicating the widespread adoption of such technologies in various industries. Therefore, preparing students in prompt engineering is crucial to meet the demands of the evolving workforce.

LLMs face challenges including the potential to generate incorrect or nonsense output (referred to as 'hallucination') and a lack of reasoning ability to solve complex problems. Therefore, improving interactions with AI-based systems is critical (White et al., 2023). Prompt engineering, also known as prompt design or prompting, involves refining inputs to generative AI (GenAI) models to produce high-quality outputs (P. Liu et al., 2023). Prompt engineering is complex, requiring trial and error, active research, and an understanding of how AI systems generate outcomes from user input (Dang et al., 2022). Currently, there is a lack of established workflows for prompt engineering, requiring extensive experimentation and evaluation by natural language processing (NLP) experts to mitigate undesirable AI outcomes (Bommasani et al., 2021; P. Liu et al., 2023). Proficiency in prompt engineering is increasingly essential for effective communication with LLM-based chatbots, highlighting its importance for students (Knoth, Tolzin, et al., 2024; White et al., 2023). Despite widespread interest in LLMs, little is known about how non-experts, such as individual learners without formal AI training through their studies, create prompts and their effectiveness in doing so (Zamfirescu-Pereira et al., 2023). This research aims to explore the prompt engineering skills of non-experts and their impact on LLM performance in a higher education context. Improving prompt engineering skills requires a thorough understanding of basic technological principles, practical experience with technology-integrated systems, and continuous skill refinement through iterative feedback loops (Meskó, 2023).

A promising instructional strategy to illustrate prompt engineering concepts is the use of worked examples (Atkinson et al., 2000; Sweller, 1988; Wittwer & Renkl, 2010). These examples provide learners with detailed task solutions and act as scaffolds to alleviate cognitive overload resulting from inefficient problem-solving methods (Janson et al., 2020; Vogel et al., 2022). They present comprehensive problem-solving approaches that are aligned with novice reasoning processes and facilitate immediate application or subsequent use through acquired frameworks (Wittwer & Renkl, 2010). Extensive research validates the superior efficacy of worked examples, particularly in early cognitive skill acquisition, compared to traditional problem-solving approaches and minimal guidance instructional methods (Kirschner et al., 2006). In addition, the integration of worked examples may augment the perceived anthropomorphism of LLMs, portraying them as assistants with agent-like qualities. This enhancement could promote greater communicative interaction and trust between students and AI-based systems engaged in prompt engineering tasks (Fink, 2012). Viewing LLMs as more anthropomorphic could encourage their agency, prompting task delegation and goal-directed strategies, and potentially ascribing persona attributes to LLMs (Baird & Maruping, 2021; Vanneste & Puranam, 2024).

At the moment, prompt engineering research predominantly emphasizes a technological perspective (Ding et al., 2021; P. Liu et al., 2023). This study attempts to introduce a skills-oriented approach to prompt engineering, recognizing its central role in providing students with the tools for effective LLM management. The guiding research question (RQ) is: *How do worked examples contribute to the acquisition of prompt engineering skills?* To answer this question, we first overview the literature on prompt engineering, worked examples, and anthropomorphism, develop hypotheses on their interplay, and then present studies to test them. The first study examines whether higher prompt quality predicts LLM output quality. The second study examines the role of worked examples for prompt engineering and LLM output quality in more detail with two different learning opportunities in the form of prompting guides. Finally, we discuss the findings, consider possible limitations, and provide implications for future research.

Theoretical Background and Hypotheses Development

Prompt Engineering

LLMs face significant challenges that require users to have the skills to use them effectively (Dwivedi et al., 2023; Zamfirescu-Pereira et al., 2023), including overcoming issues such as hallucinations and lack of common sense (Floridi & Chiriatti, 2020; Ji et al., 2023). Prompt engineering enables users to overcome these limitations and harness the potential of GenAI (P. Liu et al., 2023). The process of guiding an LLM to generate or modify text requires the creation of precise input text or instructions (White et al., 2023), which promotes bidirectional human-AI interaction for iterative prompt refinement. Studies have increasingly

turned to prompt engineering to improve the performance of generative models, reflecting the growing reliance on these technologies (Dang et al., 2022; Hou et al., 2022; P. Liu et al., 2023). Effective prompts play a critical role in defining the parameters and expectations of interactions with an LLM, dictating the structure, relevance of information, and desired output characteristics (White et al., 2023). Recognizing the importance of developing efficient prompts for LLM-based AI systems, previous research has explored the influence of prompt keywords on generative models (V. Liu & Chilton, 2021), prompt design for different tasks (Han et al., 2021), and the utility of extended context in prompts (Wu et al., 2021). Furthermore, the use of the LLM itself to refine prompts mirrors the human practice of 'thinking aloud' and provides another avenue for improving prompt design (Betz et al., 2021).

However, prompt engineering has not been studied comprehensively and systematically from a human-computer interaction (HCI) perspective, and quantitative findings within empirical research are limited. The first approaches in this area have been demonstrated in the following research projects.

Study	**Key Results**
Dang et al. (2022)	Identified challenges of prompting through focus groups, and proposed four design goals for user interfaces that support prompting.
Oppenlaender (2023)	Developed a Taxonomy of prompt modifiers for text-to-image generation.
Oppenlaender et al. (2023)	Prompt Engineering is a new type of skill for creating AI art with text-to-image generation that is non-intuitive and must be learned before it can be used.
Zamfirescu-Pereira (2023)	Non-experts explored prompts opportunistically using a prototype LLM-based chatbot tool, struggling with over-generalization and human-to-human expectations, similar to other system challenges.
This study	***Investigates how non-experts create and use prompts with an LLM and how prompt engineering skills can be trained.***

Table 1. Research Background

Oppenlaender et al. (2023) explored the potential of prompt engineering in art generation using GenAI. Their study examined the ability of untrained individuals to judge prompt quality, create and refine their own prompts, and showed that proficient prompt engineering requires practice and familiarity with relevant terminology and expressions. Building on this, Oppenlaender (2023) provided further insight into prompt engineering for text-to-image generation by presenting a taxonomy of prompt modifiers. In addition, Zamfirescu-Pereira et al. (2023) explored prompt engineering by non-experts using an LLM-based chatbot design tool, finding that while non-experts were able to generate prompt ideas, they faced challenges in systematically progressing due to limited awareness of LLM capabilities, often resorting to overly general prompts reminiscent of human-to-human instructions. Dang et al. (2022) identified barriers to prompt design through a focus group of HCI experts, highlighting issues such as lack of clear guidance in the trial-and-error process, inadequate representation of activities and outcomes, concerns about computational costs, and ethical considerations. Participants also expressed difficulties in formulating prompts tailored to specific LLM tasks. In line with this previous research (Table 1), our study investigates how non-experts create and use prompts with an LLM and how prompt engineering skills can be trained, providing insights into user practices that go beyond technology-focused approaches.

As these previous studies show, prompt engineering skills play a crucial role in shaping the output quality of LLMs. Competent prompt engineering involves the ability to produce accurate and contextually relevant input text or instructions, thereby guiding the LLM to generate or modify text effectively. Given the complex interplay between prompt quality and resulting LLM output, we suggest that individuals with advanced prompt engineering skills have the acumen to formulate prompts that elicit more coherent, accurate, and contextually appropriate responses from LLMs. Therefore, we first hypothesize (see Figure 1):

H1: Students with higher prompt engineering skills will demonstrate higher LLM output quality.

Worked Examples

Worked examples, often referred to as example-based learning, is a well-researched instructional strategy in the field of educational psychology (Schwonke et al., 2009; Wittwer & Renkl, 2010). This instructional technique provides learners with detailed solutions to tasks that serve as a support structure that minimizes the cognitive load that often results from inefficient problem-solving methods (Janson et al., 2020; Vogel et al., 2022). These examples provide a complete and accurate method for solving problems in a manner that aligns with the reasoning processes typical of novices, thereby facilitating immediate use or subsequent application through a learned framework (Wittwer & Renkl, 2010). They are especially helpful in the early stages of learning cognitive skills, over traditional problem-solving methods (Kirschner et al., 2006). This advantage is often referred to as the "worked example effect" (Sweller, 1988).

The effectiveness of using worked examples as an instructional method can be understood within the framework of Cognitive Load Theory (CLT) (Sweller, 1988). When confronted with new material, learners typically resort to generic problem-solving tactics that place a significant load on their working memory, thereby inhibiting the development of effective problem-solving schemas. Worked examples might avoid the need for learners to engage in unnecessary search efforts, reduce non-targeted prompting strategies, allow them to focus on the problem at hand, and use purposeful prompting strategies. This focus could reduce cognitive load and aid in the formation of problem-solving schemas (i.e., elaborated prompts) that are grounded in the fundamental principles of the domain. As a result, worked examples are essential for mastering knowledge that can be skillfully transferred to novel situations, thereby alleviating cognitive overload and promoting efficient learning (Atkinson et al., 2000; Sweller, 2020). Although research suggests that worked examples can be used to learn thorough self-explanations (e.g., Atkinson et al., 2000), other researchers have discussed that in unfamiliar domains (e.g., prompt engineering), learners may have difficulty correctly explaining to themselves the principles underlying the worked examples and that successful example-based learning may require additional enhancements (e.g., Berthold & Renkl, 2009; Kirschner et al., 2006; Renkl, 2002). Thus, to prevent the production of erroneous self-explanations and to enhance the example-based learning process, additional instructional explanations have been proposed that display conceptual knowledge that complements the procedural knowledge components displayed by the examples, thus developing a more complete understanding of the problem at hand. This has been shown to be particularly beneficial when students have little prior knowledge of the domain, which is insufficient to provide internal guidance (Kirschner et al., 2006; Wittwer & Renkl, 2010).

The importance of instructional explanations becomes apparent when worked examples are considered product-oriented information, as opposed to process-oriented information that explains the rationale or heuristics behind a problem in addition to the solution itself (van Gog et al., 2004). Since prompt engineering is a highly process-oriented activity, it was decided to optimize learning from worked examples by providing instructional explanations that make explicit the principles and rationale underlying the solution procedure, building knowledge about what actions to deploy, how to perform them, and why these actions work (Gott et al., 1996; van Gog et al., 2019). This instructional explanation took the form of a prompting guide (i.e., a "process worksheet") that presented both information about the underlying principles and concrete operators as examples, thus promoting the development of "far" transfer problem-solving skills (Ohlsson & Rees, 1991).

According to Ohlsson and Rees (1991), "procedures learned without conceptual understanding tend to be error-prone, easily forgotten, and do not transfer easily to novel problem types" (p. 104). We lack a clear understanding of effective prompting guides, as their appearance varies widely, with research showing different types from worked examples. As prompting is a process based primarily on procedural knowledge components, we hypothesize that students who are provided with worked examples would learn more effective prompting strategies compared to students without worked examples. We assume that worked examples work better with instructional explanation than with instructions alone. This assumption is consistent with previous research showing that worked examples containing both conceptual and procedural information lead to better performance than conceptual information alone (Kyun & Lee, 2009). This is also supported by research on cheat sheets: they are more effective when they are supplemented with additional examples (Wang et al., 2020). Similarly, a prompting guide without worked examples might still lead to some learning in the area of prompt engineering, although how to translate that knowledge into real-world application is an effort that students must make for themselves. Consequently, we propose the following hypotheses related to the effects of worked examples on prompt engineering skills (H2) and LLM output quality (H3):

H2a: Students who are provided with a worked example, enhanced by instructional explanations of how to prompt, will engage in more sophisticated prompt engineering behavior by using more purposeful prompting strategies than either students who are provided with instructions alone or no prompt engineering learning opportunity at all.

H2b: Students who are provided with an instructional explanation of how to prompt, will show more sophisticated prompting strategies, compared to a baseline condition.

H3a: Students who are provided with a worked example, enhanced by instructional explanations of how to prompt, will produce higher quality LLM output than either students who are provided with instruction alone or no prompt engineering learning opportunity at all.

H3b: Students who are provided with an instructional explanation of how to prompt, will produce LLM output of higher quality, compared to a baseline condition.

Perceived Anthropomorphism

When beginning to acquire a cognitive skill such as prompt engineering, novices have to rely on general, weak strategies (e.g., trial and error, means-ends analysis) to solve problems because they have not yet learned effective specific procedures for doing so (van Gog et al., 2019). These general, weak problem-solving strategies are not very efficient for learning because they lead to slow learning. Learners may eventually succeed in solving the problem, but as a consequence of the high working memory load, they often fail to remember which steps were actually effective. As a result, they fail to construct schemas that link problem types and effective solution procedures and remain unable to subsequently solve similar problems (Sweller & Levine, 1982; van Gog et al., 2019). This is where anthropomorphizing may help students, as attributing human characteristics to non-human entities helps to rationalize their behavior (Fink, 2012) and might lead to more effective prompt engineering. Previous studies have shown that incorporating design elements that mimic human traits and social characteristics can increase robot acceptance (Fink, 2012). This tendency stems from the human tendency to anthropomorphize objects to reduce uncertainty and establish social relationships (Epley et al., 2007). Furthermore, research suggests a preference for human-like interactions with machines (Fong et al., 2003), for instance through social features and joint action that may improve human-agent interaction (Chaves & Gerosa, 2021; Tolzin & Janson, 2023; Tolzin et al., 2023). Worked examples might play a critical role in the construction of cognitive schemas for prompt engineering and the enhancement of conceptual understanding (Sweller et al., 2011). They serve as mental models that promote analogical reasoning and lead students to perceive the LLM as knowledgeable and agentic. This is consistent with the theoretical framework proposed by Baird and Maruping (2021) for delegating tasks to agentic information system artifacts. Worked examples might not only support the understanding of prompt engineering but also might encourage students to attribute human-like characteristics to the LLM, thereby increasing its perceived communicativeness. Furthermore, well-designed worked examples that resonate with students' cognitive processes increase perceptions of the LLM's agency and reliability, leading to a higher degree of task delegation (Vanneste & Puranam, 2024). Therefore, we hypothesize:

H4: Students who are provided with a worked example, enhanced by instructional explanations of how to prompt, will perceive the LLM as more anthropomorphic than either students who are provided with instructional explanations alone or no prompt engineering learning opportunity at all.

Baird and Maruping (2021) claim that when students perceive the LLM as having human-like qualities, they become more engaged and feel empowered, leading them to delegate complex tasks to the system. This understanding of the anthropomorphic nature of the LLM shapes students' approaches to task delegation, their trust in technology, and their communication about tasks. Janson (2023) suggests that anthropomorphism fosters cognitive empathy, which may lead students to adopt intentional prompting strategies. This empathy could also foster students to attribute more agency to the LLM (Schmitt et al., 2023; Vanneste & Puranam, 2024), thus motivating purposeful prompting strategies. In addition, students who tend to view the LLM as having more human-like qualities are more likely to attribute a persona to the LLM, a strategy considered beneficial for prompting (He, 2024). Moreover, perceiving the LLM as being intelligent through anthropomorphism inspires students to engage in more complex prompt engineering (Baird & Maruping, 2021). We therefore hypothesize:

H5: The more students perceive the LLM as anthropomorphic, the more they will engage in sophisticated prompt engineering behavior by using more purposeful prompting strategies.

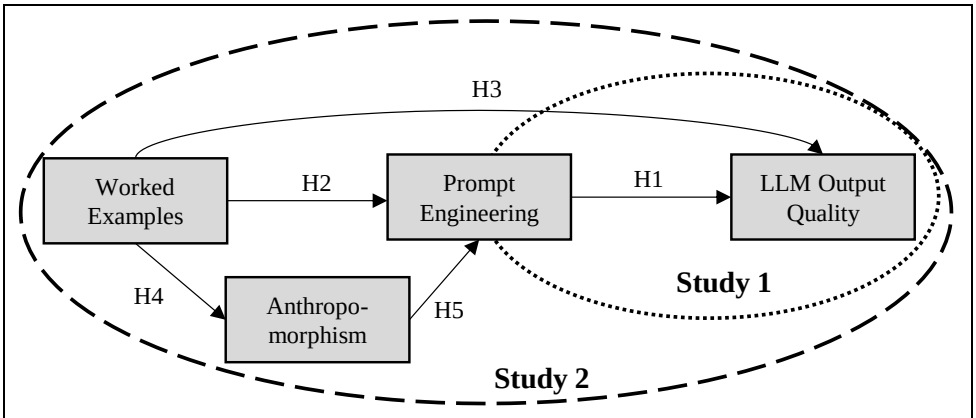

Figure 1. Conceptual Model

Study 1

Method

The first study examines the impact of prompt engineering on the quality of LLM output, while the second study examines the effectiveness of worked examples in prompt engineering and their impact on the quality of LLM outputs (Figure 1). A total of 45 undergraduate students majoring in management, psychology, or mechanical engineering were recruited for the first study and were assumed to lack expertise in AI based on their academic backgrounds. Participants were tasked with completing two tasks using a GDPR-compliant platform, which used Open AI's application programming interface (API).

To assess the prompt engineering skills of the students, two tasks were designed, each requiring a solution via an LLM. Task 1 was to produce a detailed travel plan to Andorra, which is a complex problem (Campbell, 1988) due to its multifaceted nature involving transport, accommodation, and itinerary planning. Task 2 focused on the design of a scientific project on automated essay scoring, which is also characterized by its complex nature, in line with Campbell's (1988) classification of problem tasks. Behavioral indicators were collected through structured written protocols, which recorded student-generated prompts for LLM output and the resulting LLM outputs. Following task completion, students provided reflections on the perceived ease of writing prompts, task complexity, perceived LLM output quality, and overall user experience with the GenAI. Additionally, we measured students' personal innovativeness (Agarwal & Prasad, 1998) and trust in the GenAI (Lankton et al., 2015) as controls (see Figure 3).

The quality of LLM output was assessed using an integrative complexity score, which assesses differentiation and integration. Differentiation assesses the extent to which separate aspects of the problem are considered, while integration measures the creation of complex relationships between problem features. Each LLM output was scored on a 10-point scale for both tasks (ranging from 1: minimal or no differentiation and integration, to 10: high differentiation and integration), following the method proposed by Baker-Brown et al. (1992). The first author trained a graduate student rater, and both coded the data independently and were blind to each other's coding, achieving high inter-rater reliability (IRR; Pearson correlation coefficient; Task 1: $r = .96$; $n = 42$; $p < .001$; Task 2: $r = .96$; $n = 42$; $p < .001$) as well as inter-rater agreement (IRA; weighted Cohen's kappa; Task 1: $\kappa w = 0.81$; $n = 42$; $p < 0.001$, Task 2: $\kappa w = 0.84$; $n = 42$; $p < 0.001$) (LeBreton & Senter, 2008).

In addition, prompt quality was assessed quantitatively based on six components proposed by Eager and Brunton (2023), with each component included in a prompt receiving one point: a score of 0 indicates that none of the six components were included in the prompt, while a score of 6 indicates that all prompt components were included. The components are (1) verb, (2) focus, (3) context, (4) focus and condition, (5) alignment, and (6) constraints and limitations. It is suggested that these aspects influence the quality of the outcomes produced by an LLM. The first author trained a group of graduate student raters. The data were coded independently. The coding of the prompt quality had largely good inter-rater reliability between the two raters (IRR, Pearson correlation coefficient; Task 1: $r = .83$; $n = 42$; $p < .001$; Task 2: $r = .80$; $n = 42$; p

< .001) as well as inter-rater agreement values (IRA; weighted Cohen's kappa; κw = 0.80; n = 42; p < 0.001 & κw = 0.71; n = 39; p < 0.001).

Sample

To explore how non-experts interact with LLMs and to validate hypothesis 1, an experiment was conducted in May 2023. The study cohort consisted of 45 university students aged between 19 and 35, including 15 females, 28 males and 2 participants of unspecified gender. These students were studying various disciplines, with 15 studying mechanical engineering, 6 psychology, 21 business and economics, and 3 unspecified. Notably, 28 participants had previous experience with GenAI, while 17 had no experience with such systems before the study, making prompt engineering a novel task for the latter group.

Results

From a qualitative perspective, students generally reported positive interactions with the LLM. They expressed satisfaction with the fulfillment of their expectations, the quality of the output, and the overall user experience. In addition, they indicated a willingness to use GenAI again for similar tasks and found the creation of prompts relatively easy. There was also a noticeable interest among students in using GenAI, suggesting a favorable attitude towards its adoption, which could have significant implications for future AI education initiatives, given the influential role of interest and attitude on learning success. Trust in GenAI (Lankton et al., 2015) and personal innovativeness (Agarwal & Prasad, 1998) were examined for anomalies, but no outliers were detected, and no specific hypotheses were formulated for further analysis.

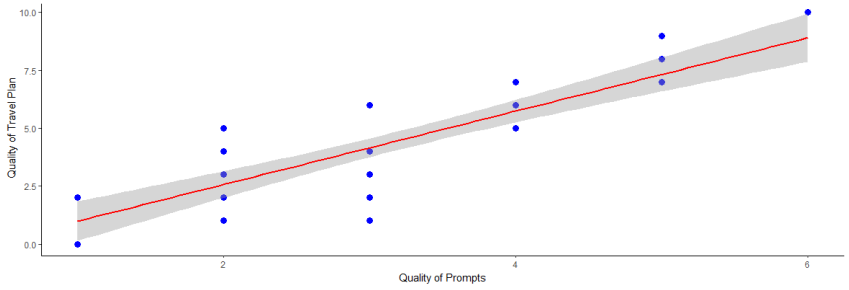

Figure 2. Impact of Prompt Engineering Quality on LLM Output (Solution of Task 1). The Blue Dots are the Datapoints

Regression analyses were conducted to test the first hypothesis (H1). The first model predicting the quality of the generated travel plan (Task 1) revealed a significant beta coefficient for prompt engineering quality on travel plan output quality (β = 1.49, t(40) = 6.78, p < .001, Figure 2). This model accounted for approximately 53% of the variance in travel plan output quality (R^2 = .535, F(1, 40) = 46.01, p < .001). Similarly, the second model predicting the quality of the three scientific project planning task solutions (Task 2) showed a significant beta coefficient (β = 1.376, t(37) = 11.502, p < .001), explaining approximately 78% of the variance in the output quality (R^2 = .782, F(1, 37) = 132.3, p < .001). Across both tasks, the association between higher quality prompt engineering behavior and higher quality LLM output was consistent, indicating that prompt engineering skills accounted for most of the variance in LLM output quality. Thus, H1 is supported.

Discussion

The results of study 1 provide initial evidence for the hypothesis 1 that a higher quality of prompts predicts the quality of LLM outputs. Prompting is a skill for producing higher quality outputs from LLMs that can potentially be learned and fostered, highlighting the importance of skill development in this area. But how can prompt engineering skills be developed and trained? It is now important to investigate what kind of educational interventions are appropriate to quickly and reliably promote the ability of students in different fields of study to construct goal-directed prompts, through a between-subjects experiment testing the

effectiveness of different instructional interventions. Therefore, in our second study, we investigated the role of worked examples for prompt engineering and the quality of LLM output in more detail (H2 to H5).

Study 2

Method

To examine the impacts of worked examples, we assessed their effectiveness within an AI-based learning environment using a between-subject experiment conducted in January 2024. This approach enables us to explore how worked examples affect both prompt engineering behaviour and the perceived anthropomorphism of the LLM. We wanted to test whether worked examples are a feasible intervention or learning opportunity to learn prompt engineering effectively and efficiently. Thus, we implemented three experimental conditions: Worked example condition (instructions + WEs) vs. instructions only vs. baseline. To provide a quick but potentially effective intervention strategy for developing prompting skills, a prompting guide was created consisting of seven prompting recommendations. The seven recommendations included (1) assigning a role to the AI, (2) priming the AI and setting the context, (3) providing structural specifications, (4) limiting the length of the AI's output, (5) providing precise descriptions of the AI's procedure and result, (6) segmenting a task and generating sequences of prompts from it, and (7) avoiding ambiguous fillers and adjectives. The worked example consisted of descriptions and examples, both good and bad, for each recommendation. For example:

Recommendation 1: Assign a role to the AI
Role assignment is a special form of priming, described in more detail in the next section. The generative AI can be assigned a specific role or 'persona' - usually at the start of a chat. This greatly influences the type, scope, and complexity of the AI's responses. Bad example: 'You are a learning assistant'. Good example: 'You are a patient and understanding university coursework tutor. Your job is to help students structure and write their essays. Your answers are helpful and motivating.

Recommendation 4: Limit the length of AI output
AI systems have a tendency to enrich their responses with additional information and to give longer answers. This is generally undesirable and fills up the AI's reminder window unnecessarily quickly when prompted. It is strongly recommended that you instruct the AI to be brief. This can be a specific number of words or a more general indication.

The initial prompting guide used was based on a compilation of existing prompting guides found on the internet and was further refined in a three-part workshop series offered by the AI-Campus, a European MOOC for learning AI (https://ki-campus.org/prompt-labor). This original form of the prompting guide, served as an experimental control condition, providing instructional explanations in the form of conceptual knowledge about prompt engineering, but not practical examples of how to use these prompting techniques. For the actual experimental, example-based learning condition, this prompting guide was adapted to include worked examples appropriate for higher education learners. The worked examples were designed following design strategies suggested in the literature, particularly the explanation of goal-operator combinations and example comparisons (Wittwer & Renkl, 2010). Therefore, each prompting recommendation found in the prompting guide additionally consisted of a positive example of how to use that strategy purposefully and a negative example of a less ideal solution. Thus, following the design principles of a worked example with instructional explanations, the intervention consisted of 1) step-by-step general instruction in the concepts and principles of prompt engineering, as research suggests that students can be helped by making subgoals in the solution procedure salient, for example, by visually isolating or labeling them (Catrambone, 1998); and 2) worked examples that are positive and negative examples of these concepts and principles, as previous research has shown that novices to a domain can benefit from comparing correct and incorrect examples (as opposed to studying only correct examples) when these are presented side by side (Durkin & Rittle-Johnson, 2012; McLaren et al., 2016). The baseline condition did not receive any prompt engineering learning material, but instead read a text about managing sustainable mass events that was the same length as the prompting guides. All conditions had a limited time frame of five minutes to complete their respective interventions. After the "learning phase" was completed, participants were automatically assigned to the problem-solving tasks to complete.

These tasks, adapted from Dell'Acqua et al. (2023), were designed to assess performance in generating ideas for new beverage products, assessing creativity, analytical skills, persuasiveness, and writing skills. The

tasks involved generating ideas for a new beverage in underserved markets and selecting the best idea while providing a rationale. The two tasks were "You are working for a beverage company in the unit developing new products. Your boss asked you to present an idea for a new product at the next manager meeting. Please, respond to the questions below: 1. Generate ideas for a new drink in markets that are underserved. Be creative and give at least 5 ideas. 2. Pick the best idea, and explain why, so that your boss and other managers can understand your thinking."

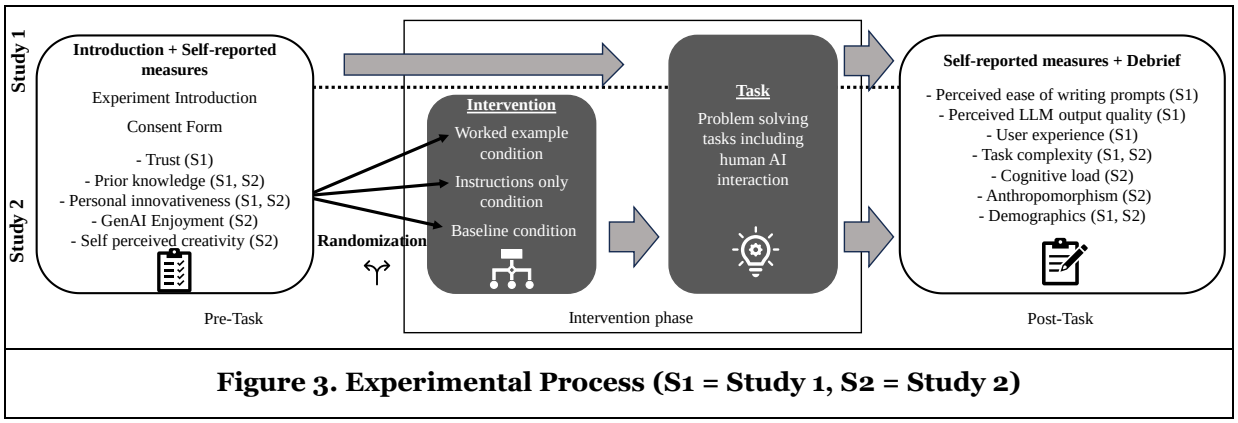

Figure 3. Experimental Process (S1 = Study 1, S2 = Study 2)

To evaluate the effectiveness of the prompts, behavioral indicators including prompt quality and LLM output quality were extracted from the chat protocols that resulted from the interaction with the LLM. The prompt quality (for tasks 1 and 2 separately) was assessed based on Eager and Brunton's (2023) six prompt components, with each component receiving one point (see study 1). Prompt scoring was conducted by two independent raters using a fully crossed rating design (Putka et al., 2008). For that, the authors trained two student teaching assistants. Any difficult or unclear cases were the subject of further discussion with the authors. To ensure that no method bias existed in our evaluation and analysis, both raters for each task were blind to the condition groups. The coding of prompt quality had largely a good inter-rater reliability between the two raters (IRR; Pearson correlation coefficient; Task 1: $r = 0.76$; $n = 227$; $p < 0.001$; Task 2: $r = 0.78$; $n = 178$; $p < 0.001$) as well as good inter-rater agreement (IRA; weighted Cohen's kappa; Task 1: $\kappa w = 0.74$; $n = 227$; $p < 0.001$, Task 2: $\kappa w = 0.78$; $n = 178$; $p < 0.001$).

In addition, anthropomorphism was measured using a six-item scale developed by Moussawi et al. (2023). The quality of the LLM output for the first task was assessed using five criteria (four specific and one general) derived from the task. Creativity describes how creative the ideas for a new drink are, while novelty describes how new and innovative the ideas are. The Context Fit criterion indicates how well the ideas for new drinks fit the context (underserved markets). The number of ideas indicates the appropriate number of ideas for the new drink (5 ideas). After scoring the four specific criteria of the LLM output, the overall output was scored. The overall impression score describes the overall impression of the generated LLM output. The quality of the LLM output for the second task was also assessed using five criteria (four specific and one general). To assess the pitches written in the second task, we adapted the approach of Carlile et al. (2018) to obtain a holistic score that captures the integrative nature of business pitch persuasion (see Figure 4). We used four specific criteria: The first criterion, persuasiveness, describes the strength and precision of the generated LLM output. Specificity assesses how detailed and specific the pitch and the associated ideas are. The third criterion, eloquence, describes how well the idea is presented, and the final criterion, evidence, reports how well the supporting statements support the business idea (new drink) and the quality of the statements. Finally, as with task 1, the overall impression score was used to assess the overall impression of the generated LLM output for task 2.

In the evaluation process, four independent raters (two for the task 1 and 2 for task 2) assessed the outputs generated by the LLM using a fully crossed rating design as outlined by Putka et al. (2008). These raters had first-hand experience with business innovation, with two of them serving as undergraduate teaching assistants and the other two being the first and second authors of the paper. The first author trained all three raters in the assessment of LLM outputs using the provided rating scales. Any difficult or unclear cases were the subject of further discussion with the authors. The formal quality of each LLM output was determined by scoring. For both tasks, all five criteria were rated individually on a scale of 1 to 5, with a score of 5 indicating the highest quality outputs. To reduce potential method bias in the scoring and

analysis, both raters who scored each task were blinded to the condition groups. Inter-rater reliability and inter-rater agreement for task 1 and 2 were assessed (IRR; Pearson correlation coefficient; Task 1: $r = 0.91$; $n = 215$; $p < 0.001$; Task 2: $r = 0.85$; $n = 166$; $p < 0.001$; IRA; weighted Cohen's kappa; Task 1: $\kappa w = 0.70$; $n = 215$; $p < 0.001$, Task 2: $\kappa w = 0.76$; $n = 166$; $p < 0.001$) indicating good reliability and inter-rater agreement across all criteria.

In addition, students completed a survey that included task complexity (according to Gupta and Bostrom, 2013), and GenAI enjoyment. Variables such as cognitive load (Krieglstein et al., 2023), prior knowledge of prompt technology, students' personal innovativeness (Agarwal & Prasad, 1998), and self-perceived creativity (Miron et al., 2004) were documented for statistical control (see Figure 3).

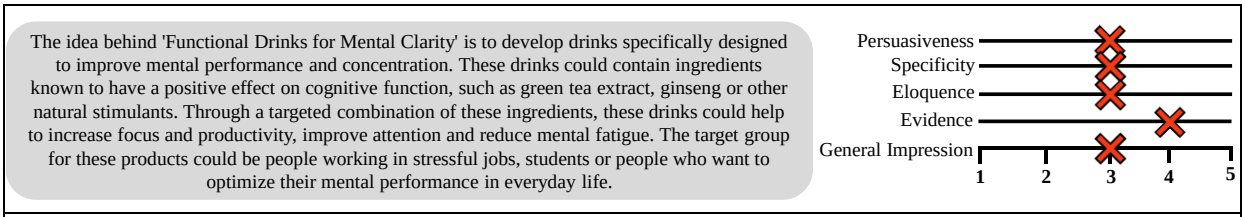

Figure 4. Example of Coding the LLM Output for Task 2. Output translated to English

Sample

This second experiment was conducted with N = 245 university students, aged between 18 and 44 years (M = 24.3; SD = 3.99). These participants identified with the following gender, n = 110 women, n = 127 men, n = 3 diverse, and n = 5 non-specified. Since the study was conducted in different subject lectures, the sample consisted of the following: Business and Administration: n = 81, Business Informatics: n = 39, Business Law: n = 52, Social Sciences: n = 39, Industrial Engineering: n = 18 and other fields like Psychology and Interdisciplinary Studies n = 16. Specifically, the classes in which the study was conducted were the following ones: A lecture on project management, two lectures on information systems, and an additional lecture on scientific methods for social sciences. Regarding previous usage of GenAI systems, n = 192 indicated that they have already used GenAI systems, n = 48 indicated that they did not use one until then and n = 5 did not know how to respond to the question. We also asked several questions to assess their prior knowledge of prompt engineering and their prior experience with prompting guides. Although most of the participants did indicate that they are new to the skill of prompt engineering (only 26.1% knew the term prompt engineering, 17.6% could explain the term, and 9.4% had read a prompting guide), almost half of them indicated that they already use specific strategies when working with AI-based systems to achieve better results (48.16%). Towards the question of "How pleasant is the use of generative artificial intelligence for you in general?", most participants indicated that they enjoy interacting with GenAI systems (M = 3.77; SD = 0.97). Following randomization, n = 85 participants were assigned to the worked example + instructional explanation condition, n = 81 participants were assigned to the instructions only condition, and n = 79 participants were assigned to the baseline condition.

Results

To test the hypotheses, a series of ANOVAs were computed (H2 - H4) and we also performed regression analyses (H1 + H5). The measures of interest considered are the intervention groups (Worked examples + Instructions vs. Instructions only vs. Baseline) and the sum scores resulting from the quantitative assessment of 1) the prompt engineering performed (quantified by the use of prompt components) and 2) the generated LLM outputs (quantified by the metrics described above). Both scores were analyzed for each given task (idea generation and pitch presentation) respectively. The analyses presented below are based on different sample sizes, as some students either did not engage in prompting activities or did not produce outputs that were suitable for the given task. Thus, only valid data points were used.

First of all, we tested if the basic effect of prompt engineering, the fact that more sophisticated prompt engineering skills produce LLM outputs of higher quality, could be replicated under conditions of higher statistical power, due to the larger sample size collected in study 2. To assess this effect, linear regressions, comparable to study 1, were performed, with the rated quality of the LLM output as criterion and rated

quality of prompt engineering as predictor. Both models yielded a significant effect of prompt engineering quality on LLM output quality (Task 1: $\beta = 1.47$, $t(201) = 4.71$, $p < 0.001$; Task 2: $\beta = 1.63$, $t(165) = 3.31$, $p = 0.001$). Therefore, H1 is confirmed, and the effect reported in study 1 could be replicated.

The main goal of this study was to investigate if prompting guides in the form of worked examples, enhanced by instructional explanations, might be a suitable educational approach to effectively foster prompt engineering skills. Consequently, we calculated ANOVA models that tested if there was a significant main effect of intervention group (Worked example + Instruction vs. Instruction vs. Baseline) on the rated quality of the prompts constructed. If significant effects were found, we calculated post-hoc t-tests to identify where the significant differences lay. The ANOVAs were again conducted for task 1 and task 2 respectively. The first model yielded a significant main effect of the experimental condition of prompt quality, $F(2, 208) = 14.95$, $p < 0.001$. The effect size, eta squared (η^2), was 0.13, indicating a moderate effect. Subsequent t-tests revealed that the worked example condition outperformed both the instructions only condition, $t(140) = 3.36$, $p < 0.001$ and the baseline condition, $t(141) = 5.30$, $p < 0.001$. Comparing the instructions only condition to the baseline, an effect could still be observed, $t(135) = 2.33$, $p = .01$. Repeating this procedure for task 2, the ANOVA model did not reveal a significant effect anymore, $F(2, 166) = 2.59$, $p = .09$. Consequently, no further t-tests were calculated. Taken together, H2a and H2b find partial support, but reasons for the absence of the effect in task 2 need to be thoroughly discussed, leaving incomplete conclusions.

Another important metric, next to the quality of prompt engineering, is the quality of the generated LLM outputs for each respective task. Thus, the statistical procedure described above was repeated with the rated quality of LLM outputs as the dependent variable. The model for the outputs of the first task again yielded a significant main effect of experimental condition on LLM output quality, $F(2, 217) = 6.63$, $p = .002$. The effect size, eta squared (η^2), was 0.06, indicating a small effect. Post-hoc t-tests revealed that the worked example condition was able to only significantly outperform the baseline condition, $t(148) = 3.56$, $p < 0.001$, but not the instructions only condition, $t(143) = 1.34$, $p = .09$. Comparing the instructions only condition to the baseline, it was able to outperform the baseline as well, $t(143) = 2.15$, $p = .02$, although descriptively on a smaller scale than the Worked example condition (see Table 2). Repeating this procedure for task 2, the ANOVA model did again reveal a significant main effect of experimental condition on generated LLM output quality, $F(2, 217) = 4.38$, $p < .05$, yielding a small effect size of eta squared (η^2) = .04. Post-hoc t-tests were in line with the findings of task 1, missing to show the ability of the worked example condition to outperform the instructions only condition, $t(143) = 0.43$, $p = .33$, but significantly outperforming the baseline condition, $t(148) = 2.70$, $p < .01$. The instructions only condition was again able to outperform the baseline condition as well, $t(143) = 2.30$, $p < .05$, although again descriptively on a smaller scale than the worked example condition (see Table 2). In conclusion, H3a finds partial support but leaves room for discussion, while H3b could be confirmed by our findings. To explain the effects of worked examples and instructional explanations, researchers typically report that cognitive load and perceived task difficulty is lower in worked example conditions, as these assist task processing and transfer. Still, within this present research neither cognitive load ($F(2, 242) = 0.11$, $p = .89$), nor perceived task difficulty ($F(2, 242) = 0.38$, $p = .68$) showed any significant differences between conditions, leaving open the question of why our intervention was able to develop its effectiveness.

Aiming to investigate the processes potentially triggered by learning with worked examples in the context of human-AI interactions with LLMs, we hypothesized that worked examples, enhanced by instructional explanations might lead to higher levels of anthropomorphism towards the LLM-based system (H4). We also expected that a rise in anthropomorphism might yield to more co-constructive task processing via more purposive prompts. For this purpose, an ANOVA model was performed, taking into account the measured level of anthropomorphism (after the intervention) as dependent variable, and experimental conditions independent variable. No significant effect could be found, $F(2, 242) = 0.76$, $p = .47$, leading to rejection of H4. Also, anthropomorphism did not yield any significant effect on prompting quality: Task 1: $b = 0.15$, $t(209) = 1.59$, $p = .11$; Task 2: $b = 0.03$, $t(167) = 0.31$, $p = .76$), leading to rejection of H5. Both of these findings point towards open research questions, of cognitive-process explanations about why the prompting guides in the form of worked examples and instructional explanations did work.

	Worked Examples + Instructional Explanations (EG1)	Instructional Explanations (EG2)	Baseline Condition (CG)
Prompt Engineering Quality (T 1)	*M* = 4.62; *SD* = 0.77	*M* = 4.12; *SD* = 0.99	*M* = 3.65; *SD* = 1.36
Prompt Engineering Quality (T 2)	*M* = 3.13; *SD* = 0.90	*M* = 2.93; *SD* = 1.11	*M* = 2.72; *SD* = 0.81
LLM Output Quality (T 1)	*M* = 15.1; *SD* = 4.36	*M* = 14.1; *SD* = 5.06	*M* = 12.1; *SD* = 5.94
LLM Output Quality (T 2)	*M* = 12.6; *SD* = 7.05	*M* = 12.1; *SD* = 6.67	*M* = 9.43; *SD* = 7.24

Table 2. Quality of Prompt Engineering and LLM Outputs (T1=Task 1, T2=Task 2)

Discussion

Discussion of Findings

The aims of our paper were to consider the prompt engineering skills of non-experts and their impact on LLM performance in a higher education context, and to reveal the role of worked examples for prompt engineering and the quality of LLM outputs. In doing so, we aim to uncover how worked examples contribute to the acquisition of prompt engineering skills.

First, in studies 1 and 2 we found empirical evidence that higher quality prompt engineering does indeed predict LLM output quality (H1). This means that we can definitely talk about prompt engineering as a quantifiable skill that differentiates between students who are able to use LLMs productively and those who may struggle to achieve their desired outcomes. This effect is strong and has been replicated. Furthermore, Oppenlaender et al. (2023) also showed that prompt engineering is a new type of skill for creating AI art with text-to-image generation that is non-intuitive and must be learned before it can be used. Following this idea, we showed in our second study that prompting guides could improve students' prompt engineering skills and the quality of their LLM output. Looking at this in more detail, we were able to show that prompting guides (worked examples enhanced by instructional explanations (H2a) as well as instructional explanations only (H2b)) improve students' prompt engineering skills in task 1. This means that both prompting guides contribute positively to students' prompt engineering skills. Furthermore, the effectiveness of these teaching methods highlights the importance of providing students with comprehensive support in understanding prompt engineering concepts and strategies.

While our results provide initial evidence that worked examples affect the ability to produce quality prompts, they were somewhat inconclusive because the effect did not show up for prompting task 2. This may be explained by the study design itself, as well as prompt engineering as an iterative process. The two tasks to be completed were sequential, meaning that task 1 had to be prompted and completed to meaningfully complete task 2. Looking more closely at the prompting strategies used by students, it becomes clear that high quality prompts were most often designed within the initial prompt, setting the stage for the conversation, using strategies such as the persona prompt, and concretizing the context and scope of the task (example prompt: *You are an innovative product developer in the beverage industry. Your role is to generate creative and marketable beverage ideas that could be successful in previously underserved markets. Your expertise in nutritional science and market trends will be fully utilized*). Once such an initial prompt was set, it did not need to be repeated for task 2; instead, the generated LLM output for task 2 likely still benefited from the quality of the prompt crafted at the beginning of the conversation (task 1). This may explain why worked example effects were found for both tasks when examining LLM outputs, whereas they were only significant for task 1 when examining prompt quality. Here we found that prompting guides (worked examples enhanced by instructional explanations (H3a) as well as instructional explanations only (H3b)) also significantly increased students' prompt engineering skills in tasks 1 and 2 compared to the baseline condition. However, here students with worked examples did not produce higher LLM outputs compared to students provided with explanations only. Thus, we can say that providing students with any kind of prompting guide helps produce higher quality LLM outputs. Furthermore, these findings highlight the potential for integrating different teaching approaches to optimize students' learning experiences in prompt engineering.

It was also interesting to note that the positive effects of the prompting guide conditions we found could not be explained by lower cognitive load, as cognitive load did not differ between groups. Thus, we could not replicate the typical worked example effect, in which a lower cognitive load facilitates processing and transfer (Wittwer & Renkl, 2010). In its revised conceptualization, cognitive load can be divided into two parts: extraneous and intrinsic load. Load that does not directly contribute to learning or problem solving is referred to as extraneous load (Krieglstein et al., 2023) and is influenced by the complexity and manner in which the task-related material is presented. Intrinsic load refers to the inherent difficulty of a task (Krieglstein et al., 2023). Both types of loads combine to determine the overall cognitive load imposed by the task and the material processed during task-solving activities. We did not distinguish between extrinsic and intrinsic load. However, when cognitive load exceeds the available capacity of working memory, the cognitive system fails to process relevant information. Applying these mechanisms to the problem-solving tasks we used in study 2, CLT in our case is concerned with managing cognitive load in such a way that there is a switch in how students invest their overall cognitive resources. This means that we do not aim to reduce overall cognitive load, but rather to reduce extraneous load and increase intrinsic load. Future research should examine extraneous and intrinsic load separately, as prompting guides may lead to better cognitive load management.

We also aimed to gain insight into the processes potentially triggered by learning with worked examples in the context of human-AI interactions with LLMs. Here, we found no support for the hypothesis that prompting guides lead to higher levels of anthropomorphism towards the LLM-based system. Furthermore, anthropomorphism had no effect on the students' prompt engineering skills. Here, it may be interesting to think about prompting strategies in a more nuanced way, as some prompting strategies (e.g., assigning a persona to the AI) may benefit more from enhanced anthropomorphism than other strategies (e.g., providing structural specifications). In future studies, the prompting strategies used by students should be analyzed in more detail to further explore the role of anthropomorphism. Nevertheless, it is worth considering why prompting guidance in the form of worked examples and instructional explanations worked more holistically, taking into account other mediators and moderators of this effect.

Contributions

Our research initially revealed a strong correlation between competent prompt engineering skills and the quality of LLM output, demonstrating that prompt engineering skills reliably predict improved LLM output. Subsequent empirical analysis underscored the central role of effective prompt engineering in fostering better LLM output, particularly for users with advanced prompt skills, thus unlocking the technology's enormous potential. In addition, our research revealed the effectiveness of prompting guides as a robust framework for cultivating prompt engineering skills. This effect is enhanced when concrete examples of prompt design are provided (i.e., the worked example effect). Furthermore, firstly, we offer a novel conceptualization of prompt engineering as an emerging skill, approached from a human-centered perspective, enriching the understanding of its importance in human-computer interaction. Second, through rigorous empirical investigation, we shed light on how worked examples influence students' ability to engage in prompt engineering behaviors, providing valuable insights into pedagogical strategies. Finally, we illuminated the role of anthropomorphism in shaping prompt engineering practices, adding depth to the discourse on how humans interact with LLMs. Given the limited number of empirical studies quantifying prompt engineering and its impact on LLM outputs, our findings represent a significant contribution to this nascent field, providing substantive empirical evidence to guide future research efforts.

Furthermore, our research adds value to educational practice by providing actionable recommendations aimed at improving students' interactions with LLMs, thus promoting more effective use of this technology in educational settings. In addition, we enrich educational practice by demonstrating how students' prompt engineering skills can be enhanced through the use of prompting guides in the form of worked examples and instructional explanations, providing a concrete way to develop skills in this area. It may not be necessary to include entire modules on prompt engineering in the curriculum. Simple prompting exercises, supported by well-designed instructional materials such as worked examples with instructional explanations, might be sufficient to improve prompting skills. The transformation of higher education towards AI might instead be better served by integrating AI literacy into the curriculum as a foundation for constructive human-AI interactions and purposeful AI use (Long & Magerko, 2020). Prompt engineering tasks can take place in the context of such a broader scope of AI education, promoting hands-on practice with AI-based systems while allowing them to be viewed in a broader perspective, building important

aspects such as knowledge, attitudes, and ethics toward AI (Knoth, Decker, et al., 2024). In addition, our study highlights the need to explore the responsible use of LLMs in education, particularly in terms of developing instructional prompts that optimize learning outcomes while maintaining ethical considerations and promoting pedagogical effectiveness. Ultimately, we are making a practical contribution by helping organizations develop learning tools that train employees for multidisciplinary roles involving AI.

Limitations, Future Research and Conclusion

As mentioned above, the study design with two tasks to be completed was sequential, meaning that task 1 had to be prompted and completed in order to meaningfully complete task 2. This was not considered in the evaluation of prompts and output. Therefore, the effect of worked examples on the ability to produce quality prompts may not be apparent. Furthermore, the operationalization of prompt engineering behavior is limited, relying exclusively on Eager and Brunton's (2023) prompt components, overlooking other potentially important aspects of prompt engineering more strongly associated with anthropomorphism. Furthermore, the lack of an objective measure of prompt engineering skills is a drawback that hinders the progress of prompt engineering research as a whole. Therefore, it is important to recommend further research into alternative approaches to modelling and assessing prompt engineering behavior.

We provide first insights into how worked examples contribute to the acquisition of prompt engineering skills in non-experts. Future research should investigate whether the effect of worked examples on the ability to produce higher quality prompts can also be found in individual learners with moderate formal AI training or in experts. In this way, it can be shown who benefits most from prompting guides and where prompting guides should be implemented most in curricula. In addition, future research should look more closely at the different tasks solved with the LLM. We only considered complex problem solving tasks, but as there is a wide variety of tasks and task complexity (Campbell, 1988) in the educational context, the effects found in our study may not apply to all tasks. Finally, it is important to note that transfer learning was required for task processing in this study, as the prompting guide was only available during a fixed learning phase. This was due to an experimental setting to control for confounding variables, and a real-world setting would not allow for the capture of actual learning. In a real-world setting, prompting guides are always present. The effects reported here may be even more pronounced if the prompting guide is present during problem solving and task processing. The present study provides initial evidence that prompting guides can be a fast and effective way to learn prompt engineering strategies, even when studied for only five minutes. Future research should investigate different scenarios, such as worked examples integrated into the LLM interface or presented at the same time.

In conclusion, the present study provides empirical evidence that worked examples, in the form of prompting guides, not only strengthen prompt engineering skills in non-experts, but also lead to higher quality outputs from LLMs. We have highlighted the growing importance of prompt engineering skill development promoting effective communication with LLM-based chatbots such as ChatGPT, underlining their relevance for students and the future workforce.

Acknowledgements

The results presented were partially developed in the research projects: Komp-HI funded by the German Federal Ministry of Education and Research (BMBF, grant 16DHBKI073) and Managing the Algorithm: Prompt Engineering for AI-based Systems as an Emerging Business Skill by the Swiss National Science Foundation (SNSF, grant number: 221281). We thank the BMBF and SNSF for supporting our research.

References

Agarwal, R., & Prasad, J. (1998). A Conceptual and Operational Definition of Personal Innovativeness in the Domain of Information Technology. *ISR*, *9*(2), 204–215. https://doi.org/10.1287/isre.9.2.204

Atkinson, R. K., Derry, S. J., Renkl, A., & Wortham, D. (2000). Learning from Examples: Instructional Principles from the Worked Examples Research. *Review of Educational Research*, *70*(2), 181–214.

Baird, A., & Maruping, L. M. (2021). The Next Generation of Research on IS Use: A Theoretical Framework of Delegation to and from Agentic IS Artifacts. *MIS Quarterly*, *45*(1), 315–341.

Baker-Brown, G., Ballard, E. J., Bluck, S., De Varies, B., Suedfeld, P., Tetlock, P. E. (1992). The conceptual/integrative complexity scoring manual. In C. P. Smith (Ed.), *Motivation and personality: Handbook of thematic content* (pp. 401–418). Cambridge University Press.

Berg, J., Raj, M., & Seamans, R. (2023). Capturing Value from Artificial Intelligence. *Acad. Manag*.

Berthold, K., & Renkl, A (2009). Instructional aids to support a conceptual understanding of multiple representations. *J. Educ. Psychol.*, *101*(1), 70–87. https://doi.org/10.1037/a0013247

Betz, G., Richardson, K., & Voigt, C. (2021). *Thinking Aloud: Dynamic Context Generation Improves Zero-Shot Reasoning Performance of GPT-2*. https://arxiv.org/pdf/2103.13033

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., Arx, S. et al. (2021). *On the Opportunities and Risks of Foundation Models*. https://arxiv.org/pdf/2108.07258

Campbell, D. J. (1988). Task Complexity: A Review and Analysis. *Acad. Manage. Rev. 13*(1), 40.

Cao, Y., Li, S., Liu, Y., Yan, Z., Dai, Y., Yu, P. S., & Sun, L. (2023). *A Comprehensive Survey of AI-Generated Content: A History of Generative AI from GAN to ChatGPT*. arxiv.org/pdf/2303.04226

Carlile, W., Gurrapadi, N., Ke, Z., & Ng, V. (2018). Give Me More Feedback: Annotating Argument Persuasiveness and Related Attributes in Student Essays. In I. Gurevych & Y. Miyao (Eds.), *Computational Linguistics (Volume 1)* (pp. 621–631), https://doi.org/10.18653/v1/P18-1058

Catrambone, R. (1998). The subgoal learning model: Creating better examples so that students can solve novel problems. *Journal of Experimental Psychology: General*.

Chaves, A. P., & Gerosa, M. A. (2021). How Should My Chatbot Interact? A Survey on Social Characteristics in Human–Chatbot Interaction Design. *Int. J. Hum–Comput. Int.*, *37*(8), 729–758.

Choi, J. H., Hickman, K. E., Monahan, A., & Schwarcz, D. B. (2023). ChatGPT Goes to Law School. *SSRN Electronic Journal*. Advance online publication. https://doi.org/10.2139/ssrn.4335905

Dang, H., Mecke, L., Lehmann, F., Goller, S., & Buschek, D. (2022). *How to Prompt? Opportunities and Challenges of Zero- and Few-Shot Learning for Human-AI Interaction in Creative Applications of Generative Models*. https://arxiv.org/pdf/2209.01390

Dell'Acqua, F., McFowland, E., Mollick, E. R., Lifshitz-Assaf, H., Kellogg, K., Rajendran, S., et al. (2023). Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality. *SSRN Electronic Journal*. doi.org/10.2139/ssrn.4573321

Ding, N., Hu, S., Zhao, W., Chen, Y., Liu, Z., Zheng, H.-T., & Sun, M. (2021). *OpenPrompt: An Open-source Framework for Prompt-learning*. https://arxiv.org/pdf/2111.01998

Durkin, K., & Rittle-Johnson, B. (2012). The effectiveness of using incorrect examples to support learning about decimal magnitude. *Learn Instr*, *22*(3), 206–214. doi.org/10.1016/j.learninstruc.2011.11.001

Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., et al. (2023). So what if ChatGPT wrote it? Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *IJIM*, *71*.

Eager, B., & Brunton, R. (2023). Prompting Higher Education Towards AI-Augmented Teaching and Learning Practice. *J. Uni. Teach 20*(5). https://doi.org/10.53761/1.20.5.02

Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On seeing human: A three-factor theory of anthropomorphism. *Psychological Review*, *114*(4), 864–886. doi.org/10.1037/0033-295X.114.4.864

Feine, J., Gnewuch, U., Morana, S., & Maedche, A. (2019). A Taxonomy of Social Cues for Conversational Agents. *Int. J. Hum. Comput. Stud.*, *132*, 138–161. doi.org/10.1016/j.ijhcs.2019.07.009

Fink, J. (2012). Anthropomorphism and Human Likeness in the Design of Robots and Human-Robot Interaction. In D. Hutchison, T. Kanade, J. Kittler, J. M. Kleinberg, F. Mattern, J. C. Mitchell, et al. (Eds.), *ICSR 2012,* (Vol. 7621, pp. 199–208), https://doi.org/10.1007/978-3-642-34103-8_20

Floridi, L., & Chiriatti, M. (2020). GPT-3: Its Nature, Scope, Limits, and Consequences. *Minds and Machines*, *30*(4), 681–694. https://doi.org/10.1007/s11023-020-09548-1

Fong, T., Nourbakhsh, I., & Dautenhahn, K. (2003). A survey of socially interactive robots. *Robotics and Autonomous Systems*, *42*(3-4), 143–166. https://doi.org/10.1016/S0921-8890(02)00372-X

Gott, S., Glaser, R., Parker Hall, E., Dibble, E., & Pokorny, R. A. (1996). A naturalistic study of transfer: Adaptive expertise in technical domains. *Transfer on Trial : Intelligence, Cognition and Instruction*. 258–288.

Gupta, S., & Bostrom, R. (2013). Research Note —An Investigation of the Appropriation of Technology-Mediated Training Methods Incorporating Enactive and Collaborative Learning. *ISR*, *24*(2), 454–469. https://doi.org/10.1287/isre.1120.0433

Han, X., Zhang, Z., Ding, N., Gu, Y., Liu, X., Huo, Y., et al. (2021). Pre-trained models: Past, present and future. *AI Open*, *2*.

He, S. (2024). *Prompting ChatGPT for Translation: A Comparative Analysis of Translation Brief and Persona Prompts*. http://arxiv.org/pdf/2403.00127

Hou, Y., Dong, H., Wang, X., Li, B., & Che, W. (2022). *MetaPrompting: Learning to Learn Better Prompts*. https://arxiv.org/pdf/2209.11486

Janson, A. (2023). How to leverage anthropomorphism for chatbot service interfaces: The interplay of communication style and personification. *CHB*, *149*, 107954. doi.org/10.1016/j.chb.2023.107954

Janson, A., Söllner, M., & Leimeister, J. M. (2020). Ladders for Learning: Is Scaffolding the Key to Teaching Problem-Solving in Technology-Mediated Learning Contexts? *Academy of Management Learning & Education*, *19*(4), 439–468. doi.org/10.5465/amle.2018.0078

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., et al. (2023). Survey of Hallucination in Natural Language Generation. *ACM Comput. Surv.*, *55*(12), 1–38.

Kirschner, P. A., Sweller, J., & Clark, R. E. (2006). Why Minimal Guidance During Instruction Does Not Work: An Analysis of the Failure of Constructivist, Discovery, Problem-Based, Experiential, and Inquiry-Based Teaching. *Educ. Psychol.*, *41*(2), 75–86. doi.org/10.1207/s15326985ep4102_1

Knoth, N., Decker, M., Laupichler, M. C., Pinski, M., Buchholtz, N., Bata, K., & Schultz, B. (2024). Developing a holistic AI literacy assessment matrix – Bridging generic, domain-specific, and ethical competencies. *Computers and Education Open*, *6*, 100177. doi.org/10.1016/j.caeo.2024.100177

Knoth, N., Tolzin, A., Janson, A., & Leimeister, J. M. (2024). AI literacy and its implications for prompt engineering strategies. *Computers and Education: Artificial Intelligence*, *6*, 100225. doi.org/10.1016/j.caeai.2024.100225

Krieglstein, F., Beege, M., Rey, G. D., Sanchez-Stockhammer, C., & Schneider, S. (2023). Development and Validation of a Theory-Based Questionnaire to Measure Different Types of Cognitive Load. *Educational Psychology Review*, *35*(1). doi.org/10.1007/s10648-023-09738-0

Kyun, S. A., & Lee, H. (2009). The effects of worked examples in computer-based instruction: Focus on the presentation format of worked examples and prior knowledge of learners. *Asia Pacific Education Review*, *10*(4), 495–503. doi.org/10.1007/s12564-009-9044-x

Lankton, N., McKnight, D. H., & Tripp, J. (2015). Technology, Humanness, and Trust: Rethinking Trust in Technology. *JAIS*, *16*(10), 880–918. doi.org/10.17705/1jais.00411

LeBreton, J. M., & Senter, J. L. (2008). Answers to 20 Questions About Interrater Reliability and Interrater Agreement. *Organ. Res. Methods*, *11*(4), 815–852. doi.org/10.1177/1094428106296642

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2023). Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. *ACM Comput. Surv.*, *55*(9), 1–35.

Liu, V., & Chilton, L. B. (2021). *Design Guidelines for Prompt Engineering Text-to-Image Generative Models*. https://arxiv.org/pdf/2109.06977

Long, D., & Magerko, B. (2020). What is AI Literacy? Competencies and Design Considerations. In R. Bernhaupt, F., Mueller, D. Verweij, J. Andres, J. McGrenere, A. Cockburn, et al. (Eds.), *CHI 2020 Proceedings* (pp. 1–16). ACM.

McLaren, B. M., van Gog, T., Ganoe, C., Karabinos, M., & Yaron, D. (2016). The efficiency of worked examples compared to erroneous examples, tutored problem solving, and problem solving in computer-based learning environments. *CHB*, *55*, 87–99. doi.org/10.1016/j.chb.2015.08.038

McLean, G., & Osei-Frimpong, K. (2019). Hey Alexa … examine the variables influencing the use of artificial intelligent in-home voice assistants. *CHB*, *99*, 28–37. doi.org/10.1016/j.chb.2019.05.009

Meskó, B. (2023). Prompt Engineering as an Important Emerging Skill for Medical Professionals: Tutorial. *Journal of Medical Internet Research*, *25*, e50638. doi.org/10.2196/50638

Miron, E., Erez, M., & Naveh, E. (2004). Do personal characteristics and cultural values that promote innovation, quality, and efficiency compete or complement each other? *Journal of Organizational Behavior*, *25*(2), 175–199. doi.org/10.1002/job.237

Moussawi, S., Koufaris, M., & Benbunan-Fich, R. (2023). The role of user perceptions of intelligence, anthropomorphism, and self-extension on continuance of use of personal intelligent agents. *EJIS*, *32*(3), 601–622. doi.org/10.1080/0960085X.2021.2018365

Nguyen, H. (2023). Role design considerations of conversational agents to facilitate discussion and systems thinking. *Computers & Education*, *192*, 104661. doi.org/10.1016/j.compedu.2022.104661

Ohlsson, S., & Rees, E. (1991). The Function of Conceptual Understanding in the Learning of Arithmetic Procedures. *Cognition and Instruction*, *8*(2), 103–179. doi.org/10.1207/s1532690xci0802_1

Oppenlaender, J. (2023). *A Taxonomy of Prompt Modifiers for Text-To-Image Generation*. https://arxiv.org/pdf/2204.13988

Oppenlaender, J., Linder, R., & Silvennoinen, J. (2023). *Prompting AI Art: An Investigation into the Creative Skill of Prompt Engineering*. https://arxiv.org/pdf/2303.13534

Putka, D. J., Le, H., McCloy, R. A., & Diaz, T. (2008). Ill-structured measurement designs in organizational research: Implications for estimating interrater reliability. *The Journal of Applied Psychology*, *93*(5), 959–981. doi.org/10.1037/0021-9010.93.5.959

Rahman, M. M., & Watanobe, Y. (2023). ChatGPT for Education and Research: Opportunities, Threats, and Strategies. *Applied Sciences*, *13*(9), 5783. doi.org/10.3390/app13095783

Rasul, T., Nair, S., Kalendra, D., Robin, M., Oliveira Santini, F. de, Ladeira, W. J.et al. (2023). The role of ChatGPT in higher education: Benefits, challenges, and future research directions. *JALT* (6(1)).

Renkl, A. (2002). Worked-out examples: instructional explanations support learning by self-explanations. *Learning and Instruction*, *12*(5), 529–556. doi.org/10.1016/S0959-4752(01)00030-5

Rudolph, J., Tan, S., & Tan, S. (2023). War of the chatbots: Bard, Bing Chat, ChatGPT, Ernie and beyond. The new AI gold rush and its impact on higher education. *JALT 6*(1). doi.org/10.37074/jalt.2023.6.1.23

Schmitt, A., Zierau, N., Janson, A., & Leimeister, J. M. (2023). The Role of AI-Based Artifacts' Voice Capabilities for Agency Attribution. *JAIS*, *24*(4), 980–1004. doi.org/10.17705/1jais.00827

Schwonke, R., Renkl, A., Krieg, C., Wittwer, J., Aleven, V., & Salden, R. (2009). The worked-example effect: Not an artefact of lousy control conditions. *CHB*, *25*(2), 258–266. doi.org/10.1016/j.chb.2008.12.011

Sok, S., & Heng, K. (2023). ChatGPT for Education and Research: A Review of Benefits and Risks. *SSRN Electronic Journal*. Advance online publication. doi.org/10.2139/ssrn.4378735

Sweller, J. (1988). Cognitive Load During Problem Solving: Effects on Learning. *Cognitive Science*, *12*(2), 257–285. doi.org/10.1207/s15516709cog1202_4

Sweller, J. (2020). Cognitive load theory and educational technology. *Educational Technology Research and Development*, *68*(1), 1–16. doi.org/10.1007/s11423-019-09701-3

Sweller, J., Ayres, P., & Kalyuga, S. (2011). *Cognitive Load Theory*. *Explorations in the learning sciences, instructional systems and performance technologies*. doi.org/10.1007/978-1-4419-8126-4

Sweller, J., & Levine, M. (1982). Effects of goal specificity on means–ends analysis and learning. *JEP:LMC*, *8*(5), 463–474. doi.org/10.1037/0278-7393.8.5.463

Tolzin, A., & Janson, A. (2023). Mechanisms of Common Ground in Human-Agent Interaction: A Systematic Review of Conversational Agent Research. *HICSS 2023*.

Tolzin, A., Körner, A., Dickhaut, E., Janson, A., Rummer, R., & Leimeister, J. M. (2023). Designing Pedagogical Conversational Agents for Achieving Common Ground. DESRIST 2023 (Vol. 13873, pp. 345–359). doi.org/10.1007/978-3-031-32808-4_22

van Gog, T., Paas, F., & van Merriënboer, J. J. (2004). Process-Oriented Worked Examples: Improving Transfer Performance Through Enhanced Understanding. *Instructional Science*, *32*(1/2), 83–98.

van Gog, T., Rummel, N., & Renkl, A. (2019). Learning How to Solve Problems by Studying Examples. In *The Cambridge Handbook of Cognition and Education* (pp. 183–208). Cambridge University Press.

Vanneste, B., & Puranam, P. (2024). Artificial Intelligence, Trust, and Perceptions of Agency. *Academy of Management Review*, *15*(4), 571. doi.org/10.2139/ssrn.3897704

Vogel, F., Kollar, I., Fischer, F., Reiss, K., & Ufer, S. (2022). Adaptable scaffolding of mathematical argumentation skills: The role of self-regulation when scaffolded with CSCL scripts and heuristic worked examples. *ijCSCL*, *17*(1), 39–64. doi.org/10.1007/s11412-022-09363-z

Wang, Z., Sundin, L., Murray-Rust, D., & Bach, B. (2020). Cheat Sheets for Data Visualization Techniques. In R. Bernhaupt, F. Mueller, D. Verweij, J. Andres, J. McGrenere, A. Cockburn, et al. (Eds.), *CHI 2020* (pp. 1–13). ACM. doi.org/10.1145/3313831.3376271

White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., et al. (2023). *A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT*. http://arxiv.org/pdf/2302.11382v1

Wittwer, J., & Renkl, A. (2010). How Effective are Instructional Explanations in Example-Based Learning? A Meta-Analytic Review. *Educational Psychology Review*, *22*(4), 393–409.

Wu, T., Terry, M., & Cai, C. J. (2021). *AI Chains: Transparent and Controllable Human-AI Interaction by Chaining Large Language Model Prompts*. https://arxiv.org/pdf/2110.01691

Zamfirescu-Pereira, J. D., Wong, R. Y., Hartmann, B., & Yang, Q. (2023). Why Johnny Can't Prompt: How Non-AI Experts Try (and Fail) to Design LLM Prompts. In *Proceedings CHI 2023* (pp. 1–21).

Zhu, C., Sun, M., Luo, J., Li, T., & Wang, M. (2023). How to harness the potential of ChatGPT in education? *Knowledge Management & E-Learning: An International Journal*(15(2)), 133–152. https://doi.org/10.34105/j.kmel.2023.15.008